Swin-Unet: Unet-like Pure Transformer for Brain Tumor Segmentation

Omar Ayman, Muhammed Essam, Ahmed Abdelgawad, Yahya Mohamed, Eng. Ahmed Abdelsalam, Dr. Omar Fahmy
Communications and Information Engineering
University of Science and Technology, Zewail City of Science, Technology, and Innovation
Email: s-omar.mahmoud@zewailcity.edu.eg, s-mo.essam@zewailcity.edu.eg, s-ahmed.abdelgawad@zewailcity.edu.eg,
s-yahya.mohamed@zewailcity.edu.eg
ahmed.abdelsalam@zewailcity.edu.eg, ofahmy@zewailcity.edu.eg

***Abstract*—In recent years, Computer Vision community has made significant strides in medical image segmentation, particularly through the use of convolutional neural networks (CNNs). However, CNN-based methods have limitations in capturing long-range dependencies due to the inherent locality of convolution operations. Inspired by the success of Transformers in natural language processing, we introduce Swin-Unet, a novel model for 2D medical image segmentation that leverages the Swin Transformer. Swin-Unet is the first pure Transformer-based U-shaped architecture, featuring an encoder, bottleneck, decoder, and skip connections built on Swin Transformer blocks. The model processes input images by splitting them into non-overlapping patches, treating each patch as a token. These tokens are then processed by the Transformer-based encoder to learn deep feature representations. The decoder up-samples these features using a novel patch-expanding layer and integrates multi-scale features from the encoder via skip connections to restore spatial resolution for precise segmentation. Extensive experiments on brain tumor datasets demonstrate that Swin-Unet achieves high accuracy and robust generalization. We used Binary cross-entropy loss to assess the performance of our model, which scored 98.9% accuracy. Swin-Unet represents a significant advancement in leveraging Transformer architectures for medical image segmentation, addressing CNN limitations, and opening new research avenues.**

Keywords— Medical Image Segmentation; Swin Transformer; Deep Learning; U-Net; Encoder-Decoder Architecture; Transformer-based Model

I. Introduction

With the rapid advancement of Deep Learning, Computer Vision has increasingly been applied to medical image analysis. Image segmentation, in particular, is a crucial component of medical image analysis as it underpins computer-aided diagnosis and image-guided clinical surgery.

Traditional methods for medical image segmentation have predominantly relied on fully convolutional neural networks (FCNNs) with a U-shaped architecture. A quintessential example is the U-Net, which features a symmetric Encoder-Decoder structure connected by skip connections. In this architecture, the encoder extracts deep features through a series of convolutional layers and down-sampling operations, while the decoder up-samples these features to the original image resolution for precise pixel-level prediction. The skip connections facilitate the fusion of high-resolution features from the encoder with the decoder, thereby mitigating the loss of spatial information due to down-sampling. This structural design has proven highly effective, leading to the development of various U-Net variants such as 3D U-Net, Res-UNet, U-Net++, and UNet3+.

Despite their success, CNN-based methods have inherent limitations in capturing long-range dependencies and global context due to the locality of convolution operations. Several approaches have attempted to address these limitations using techniques such as Atrous convolution, self-attention mechanisms, and image pyramids. However, these methods still struggle to model long-range dependencies effectively.

Inspired by the success of Transformers in natural language processing (NLP), researchers have begun exploring their application in the vision domain. The Vision Transformer (ViT) demonstrated that Transformers could achieve competitive performance in image recognition by treating 2D image patches as input tokens and leveraging large-scale pre-training datasets. Further advancements like the Data-efficient Image Transformer (DeiT) showed that robust Transformer models could be trained on mid-sized datasets. Building on these developments, the Swin Transformer introduced a hierarchical architecture with shifted windows, achieving state-of-the-art results in image classification, object detection, and semantic segmentation.

Motivated by these advancements, we propose Swin-Unet, a novel Transformer-based model for 2D medical image segmentation. To our knowledge, Swin-Unet is the first pure Transformer-based U-shaped architecture incorporating an encoder, bottleneck, decoder, and skip connections, all built upon the Swin Transformer blocks. This architecture processes input medical images by splitting them into non-overlapping patches, treating each patch as a token. These tokens are then fed into the Transformer-based encoder to learn deep feature representations. The context features are subsequently up-sampled by the decoder with a patch-expanding layer and fused with multi-scale features from the encoder via skip connections. This process restores the spatial resolution of feature maps, enabling precise segmentation predictions.

Our experiments on brain tumor Segmentation datasets demonstrate that Swin-Unet achieves excellent segmentation

accuracy and robust generalization. Our main contributions can be summarized as follows: 1. We use Swin-Unet with brain tumor segmentation where most of the previously published papers use U-net architecture. 2. Employed the Swin-Unet architecture with LGG Segmentation Dataset rather than the BraTs datasets, which are widely used in the literature. 3. Implementing a Swin-Unet architecture tailored to the dataset used.

This paper represents Swin-Unet represents a significant step forward in leveraging Transformer architectures for medical image segmentation, addressing the limitations of CNN-based methods and opening new avenues for research and clinical applications.

II. Related work

Over the past five years, numerous innovative and effective architectures have been developed, significantly advancing the field of medical image segmentation. This period has seen a rapid evolution of methodologies, from traditional approaches to cutting-edge deep learning models.

A. CNN based methods

Early methods for medical image segmentation primarily relied on contour-based and traditional machine learning algorithms. With the advancement of deep convolutional neural networks (CNNs), the introduction of U-Net revolutionized medical image segmentation due to its U-shaped architecture that allows for precise localization while maintaining contextual information. Variants of U-Net, such as Res-UNet, Dense-UNet, U-Net++, and UNet3+, have been developed to enhance segmentation accuracy and robustness. These models have also been extended to three-dimensional data, leading to architectures like 3D-Unet and V-Net. The success of CNN-based methods in medical image segmentation can be attributed to their powerful representation and learning capabilities.

B. Vision transformers

Transformers were first introduced for natural language processing tasks, particularly machine translation. Their success in NLP has driven researchers to explore their application in computer vision. The Vision Transformer (ViT) marked a significant milestone by demonstrating competitive performance in image recognition tasks. However, ViT's requirement for extensive pre-training on large datasets posed challenges. The introduction of DeiT (Data-efficient Image Transformers) addressed these issues by proposing training strategies that enable effective training on smaller datasets like ImageNet. The hierarchical vision Transformer, known as Swin Transformer, further advanced the field by introducing a shifted windows mechanism, achieving state-of-the-art results in various vision tasks, including image classification, object detection, and semantic segmentation. In this study, we leverage the Swin Transformer to build a U-shaped Encoder-Decoder architecture with skip connections, aimed at enhancing medical image segmentation performance.

C. Self-Attention and Transformers to Complement CNNs

In recent years, the self-attention mechanism, a key component of Transformers, has been integrated into CNN architectures to improve their performance. Additive attention gates have been incorporated into U-shaped architectures for medical image segmentation. Despite these advancements, such approaches remain primarily CNN-based. Efforts to combine Transformers with CNNs aim to leverage the strengths of both architectures, breaking CNNs' dominance in medical image segmentation. For instance, a hybrid model that combines Transformers with CNNs has shown promise in enhancing 2D medical image segmentation. These hybrid models have been applied to multimodal brain tumor segmentation and 3D medical image segmentation, demonstrating improved segmentation capabilities. Unlike these hybrid approaches, our work focuses on exploring the application potential of pure Transformer architectures in medical image segmentation.

D. CNN and Variants

CNNs have been a cornerstone in computer vision, with architectures like AlexNet, VGG, GoogleNet, ResNet, DenseNet, HRNet, and EfficientNet driving significant advancements. Innovations such as depthwise convolution and deformable convolution have further refined CNN performance. While CNNs remain the primary backbone for many applications, the potential of Transformer-like architectures for unified modeling between vision and language is increasingly recognized. Our study aims to contribute to this evolving paradigm by demonstrating the effectiveness of a pure Transformer-based model for medical image segmentation.

E. Self-Attention-Based Backbone Architectures

Inspired by the success of self-attention layers and Transformer architectures in NLP, several works have explored replacing spatial convolution layers in popular architectures like ResNet with self-attention layers. These modifications achieve better accuracy/FLOPs trade-offs, albeit with increased memory access costs. To address this, we propose shifting windows between consecutive layers to enhance memory efficiency and overall performance.

F. Transforming Vision Tasks with Transformers

The pioneering work on Vision Transformers (ViT) directly applied Transformer architectures to image patches, achieving impressive results in image classification. Despite the need for large-scale datasets for training, advancements like DeiT have made ViTs more accessible for smaller datasets. The Swin Transformer further enhances the applicability of Transformers to dense vision tasks, offering a more efficient architecture that scales linearly with image size.

These different architectures show that the field of medical image segmentation has seen significant advancements with the introduction and evolution of CNN-based methods. However, the integration of Transformer architectures, particularly with models like the Swin Transformer, represents a promising direction for further improvements. By leveraging the strengths

of both CNNs and Transformers, our work aims to push the boundaries of medical image segmentation, providing a robust benchmark for future research.

III. Dataset

Our system is trained and analyzed with the **LGG Segmentation Dataset**, available on Kaggle [1]. This dataset was used in the study "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm" by Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski.

A. Patient Population

The data was obtained from The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). We identified 120 patients with preoperative imaging data, excluding 10 due to missing genomic information, resulting in 110 patients from five institutions. The patients were split into 22 subsets for cross-validation.

B. Imaging data

Imaging data from TCIA included multiple modalities, with FLAIR used if others were missing. The dataset included 101 patients with all sequences, 9 missing post-contrast, and 6 missing pre-contrast sequences. Preoperative data was analyzed, and FLAIR images were manually annotated and verified by radiologists.

C. Genomic Data

Genomic data included DNA methylation, gene expression, DNA copy number, and microRNA expression, focusing on six molecular classifications correlated with tumor shape features:

1) IDH mutation and 1p/19q co-deletion
2) RNASeq clusters
3) DNA methylation clusters
4) DNA copy number clusters
5) microRNA expression clusters
6) Cluster of clusters

IV. Methodology

A. Data Preprocessing

Data preprocessing is a critical step in preparing the input for efficient and effective deep learning model training. The preprocessing pipeline includes the following steps:

- **Conversion to PNG**: All images and their masks were converted from .tif extension to .png extension. This was done to ease the entire process of preprocessing.
- **Image Resizing**: All images are resized to a standard dimension (224x224 pixels) to ensure uniformity and to reduce computational load.
- **Grayscale Conversion**: Masks were converted to grayscale to reduce the number of channels processed by the neural networks, thus simplifying the model's complexity without significant loss of information.
- **Data Augmentation**: To improve model robustness and generalizability, data augmentation techniques such as rotation, scaling, and horizontal flipping are applied.

B. Model Development

We have explored the use of the Swin-Unet architecture for the task of brain tumor segmentation. The Swin-Unet combines the strengths of the Swin Transformer and the well-established U-net architecture. This hybrid approach leverages the Swin Transformer's ability to effectively capture long-range dependencies and spatial information while maintaining U-net's prowess in preserving local details and producing segmentation maps with high spatial resolution. The multi-scale feature representation and the self-attention mechanisms employed by the Swin-U-net allow the model to holistically understand the complex structure of brain tumors, leading to improved segmentation accuracy and robustness. We have carefully tailored the Swin-Unet architecture to the specific requirements of brain tumor segmentation, incorporating domain-specific enhancements and optimization techniques to further enhance its performance on this clinically vital task. The results of our extensive evaluations have been highly promising, demonstrating the Swin-Unet's ability to outperform traditional Unet-based models and set new benchmarks in this important field of medical image analysis.

- **Model Architecture**: The following figure shows the architecture of the used Swin-Unet as described before.

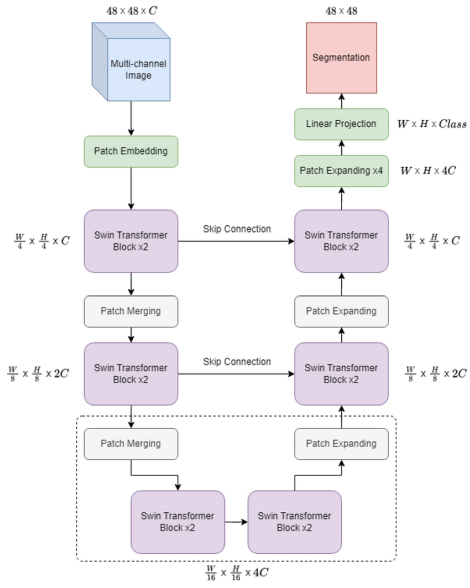

Fig. 1. Swin-Unet architecture

- **Training**: For training the Swin-Unet model, we employed the Adam optimization algorithm with a learning rate of 0.001. The model was trained to minimize a binary cross entropy loss function, which is well-suited for the binary segmentation task of differentiating between tumor and non-tumor regions. By leveraging the adaptive gradients and momentum properties of the Adam optimizer, the training process was able to efficiently converge towards the optimal model parameters that best captured the intricate patterns and boundaries of brain tumors in the input medical images. The binary cross

entropy loss function provided a robust learning signal, encouraging the model to accurately classify each pixel as either belonging to the tumor or the background. This combination of the Adam optimizer and the binary cross entropy loss enabled the Swin-Unet architecture to be trained effectively for the brain tumor segmentation task.

- **Hardware**: We leveraged the hardware resources provided by the Kaggle platform. Specifically, we utilized the GPU-powered Kaggle Notebooks, which offered access to high-performance NVIDIA GPU P100 accelerators.

V. Results

The results highlight the effectiveness of the Swin-Unet architecture in brain tumor segmentation, which outperforms the traditional CNN architectures used for image segmentation.

A. Model Training and Validation

- **Accuracy and Loss Metrics**: Our Swin-Unet architecture used binary-cross entropy loss function to segment the tumors. Its formula is given by:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

Our model achieved a validation accuracy of up to 98.9%, outperforming most of the known architectures. The following table shows some traditional CNN architectures and ViTs used on various brain tumor segmentation datasets.

Model	Accuracy
ViT-based DNN	97.98%
QFS-Net	98.23%
Swin-Unet	98.9%

TABLE I. Accuracies of different architectures

The next figure shows three different MRI images of the brain, with their masks and the prediction for each mask. If the mask is empty, then the corresponding MRI scan doesn't show any tumors.

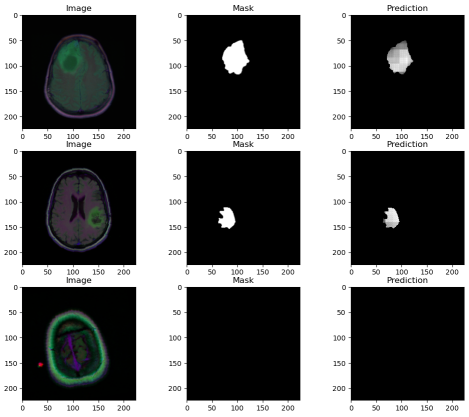

Fig. 2. Predictions of the Swin-Unet

VI. Conclusion

In conclusion, the Swin-Unet model has demonstrated its effectiveness as a powerful tool for the segmentation of brain tumors in medical images. By seamlessly integrating the Swin Transformer's ability to capture long-range dependencies with the Unet's preservation of local spatial details, the model was able to accurately delineate the complex boundaries and structures of brain tumors. The training process, which leveraged the Adam optimizer and binary cross entropy loss on GPU-accelerated Kaggle hardware, enabled efficient convergence towards an optimal set of model parameters. The results of this project showcase the Swin-Unet's potential to assist medical professionals in the early detection and diagnosis of brain cancer, potentially leading to improved patient outcomes. Going forward, further research could explore the model's generalization capabilities on diverse brain tumor datasets, as well as its applicability to other medical image segmentation tasks. Overall, this work highlights the valuable role that advanced deep learning architectures can play in advancing the field of computer-aided diagnosis and analysis of complex medical imagery.

A. Future Work

Future work on the area of brain tumor segmentation should focus on:

- **Dataset Expansion**: Expanding the dataset to include a broader range MRI scans of different modalities.
- **Swin transformers**: Experimenting Swin transformers as encoders with different CNN architectures used for medical image segmentation.

Acknowledgment

The authors would like to thank Dr. Omar Fahmy and Eng. Ahmed Abdelsalam for their guidance and support throughout this project.

References

[1] kaggle.com - Brain MRI segmentation [https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation].

[2] [Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. (2021, May 12). Swin-UNet: UNET-like pure transformer for medical image segmentation. arXiv.org. https://arxiv.org/abs/2105.05537

[3] Buda, M., Saha, A., Mazurowski, M. A. (2019). Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Computers in Biology and Medicine, 109, 218–225. https://doi.org/10.1016/j.compbiomed.2019.05.002

[4] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021, March 25). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv.org. https://arxiv.org/abs/2103.14030

[5] Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y., Zhou, H. (2022). Deep learning based brain tumor segmentation: a survey. Complex Intelligent Systems, 9(1), 1001–1026. https://doi.org/10.1007/s40747-022-00815-5

[6] Zheng, P., Zhu, X., Guo, W. (2022). Brain tumour segmentation based on an improved U-Net. BMC Medical Imaging, 22(1). https://doi.org/10.1186/s12880-022-00931-1

[7] Mostafa, A. M., Zakariah, M., Aldakheel, E. A. (2023). Brain tumor segmentation using deep learning on MRI images. Diagnostics, 13(9), 1562. https://doi.org/10.3390/diagnostics13091562

[8] Gupta, A., Dixit, M., Mishra, V. K., Singh, A., Dayal, A. (2023, April 29). Brain Tumor Segmentation from MRI Images using Deep Learning Techniques. arXiv.org. https://arxiv.org/abs/2305.00257

[9] Montaha, S., Azam, S., Rafid, A. K. M. R. H., Hasan, M. Z., Karim, A. (2023). Brain Tumor Segmentation from 3D MRI Scans Using U-Net. SN Computer Science/SN Computer Science, 4(4). https://doi.org/10.1007/s42979-023-01854-6